MIE 1624 Assignment 1

Zorina Song

Background and Dataset

The data was from "Kaggle ML & DS Survey Challenge" which provided 25793 participants with 369 features, but only the information about Age, Gender, Country, Education, Job, Experience, and Salary of 15391 participants were considered as important features in this analysis for the purpose of understanding women's representation in Data Science and Machine Learning, and the effect of education on income level.

Part 1. Analyze the dataset and summarize the main characteristics.

- **1.1 Salary & Gender**
  For comparing the amount of each job taken by men and women, the horizontal bar plots was shown in *Figure 1.1*. Obviously, the spreads of women and men on choosing each job seemed similar, as the Data Scientists was the primary group, and Developer Relations/Advocacy took the least amount. However, the proportions of male employees on each job position was performed overwhelmingly at the same time, and only an approximate of 25% or less were taken by women in each position.
- **1.2 Highest Level of Education & Salary**
  Generally, the education could play an important role on skill-gaining which also affect the salary positively. With the trend lines in *Figure 1.2*, the median and mean salary regarding the education levels, from lower to higher, were displayed similarly, in spite of the observation that the mean kept much higher than the median for each education level, and the group with profession doctorate got lower median salary than the ones with Master's degree. Apart from that, the group with Bachelor's degree had a lower median and mean salary comparing with the group took some college/ university study without earning a Bachelor's degree.
- **1.3 Experience/Age & Salary**
  Pay may go up because of the experience comes with age. The *Figure 1.3* indicated the upward trends of both mean and median salary as the groups got more years of coding experience. However, according to *Figure 1.4,* the mean and median salary of the groups with 60-year or higher old might not follow the tendency, which meant the salary of the one around retirement age were not much predictable.

Pro and Con: The comparison of the trends were clear, but neither mean nor median was representative enough.

Part 2. Estimate the difference between average salary of men and women.

The descriptive analysis of *Table 2.0* suggested a huge gap between the salary of men and women, except the same extreme values. For a further step to compare the mean average salary of the whole population by gender, uncertainty in the survey data was assumed, then a two-sample t test with 0.05 threshold was conducted. As the p-value was 8.08881e-15 which was less than the threshold, the higher mean average salary for male

employees observed before was most likely not due to change, and the difference was statistically significant.
By taking a random sample with replacement from the data used for both men and women, with sample size as the original 12642 and 2482 respectively, the bootstrap was introduced and repeated 1000 times, as the mean was not representative sometimes. The distributions of the bootstrapped data and their difference were performed by *Figure 2.1* and *Figure 2.2,* which were roughly normal, and suggested the mean pay in jobs related with coding was approximately $14,000 to $18,000 higher for men than women in most cases. Additionally, the two-sample t test with 0.05 threshold on the bootstrapped data resulted in a p-value of 0, showing the difference was statistically significant, so we concluded that the mean salary of men tended to be higher than that of women obviously in the jobs related with coding.
Pro and Con: The t-test could be used when the population parameters were not know, but it only worked for determining whether the difference exist as we assumed, but fail to show how they differed. The bootstrap was straightforward to derive the estimates of confidence intervals, but the intervals may be not unreliable or the sample of 15391 participants attending the competition was not representative.

Part 3. Estimate the difference between average salary of the groups with different highest level of education, mainly Bachelor's degree, Master's degree, and Doctoral degree.

The descriptive statistics from *Table 3.0 had* showed most employees were with Master's degree and then Bachelors' degree, and the salaries at the mean and all the percentiles would increase from Bachelor's to Doctoral degree. The ANOVA with a p-value of 5.1077e-48 indicating the significance and also suggested the mean salary would rise with improved qualification in education.
Next, the data was also bootstrapped like above but for the groups with different levels of education. The distributions of the bootstrapped data and their difference roughly followed the normal distribution, as shown by *Figure 3.1 and 3.2*. The mean salary would be mainly $15,000 to $20,000 higher for the group with Doctoral degree rather than Master's degree, the difference even enlarged to a range of $32,000 to $38,000 mostly compared with Bachelor's degree. With the p-value of 0 from the ANOVA test on the bootstrapped data indicating the statistical significance, we concluded that the difference was quite statistically significant, and the mean average would rise if the academic degree obtained was higher.
Pro and Con: Similarly as above. (Anova also failed to show how the mean differed)

Conclusion

The amounts of female employed in the jobs related with Data Science and Machine Learning were quite lower than those of male, and they also earning less regarding of the mean average. Besides, for the academic degrees of Bachelor, Master and Doctoral, a higher education degree could result in an increase of the mean salary for these groups.

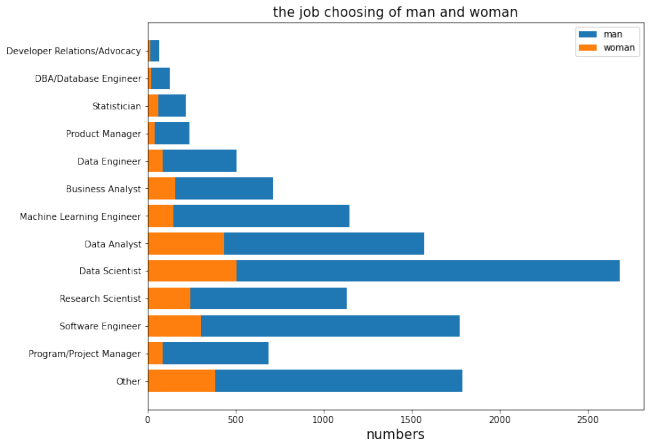

Figure 1.1

Gender	count	mean	std	min	25%	50%	75%	max
Man	12642.0	51193.600696	99979.274378	1000.0	2000.0	20000.0	60000.0	1000000.0
Woman	2482.0	34816.881547	72017.347888	1000.0	1000.0	7500.0	50000.0	1000000.0

Table 2.0

Education	count	mean	std	min	25%	50%	75%	max
Bachelor's degree	4777.0	35578.291815	89382.060777	1000.0	1000.0	7500.0	40000.0	1000000.0
Master's degree	6799.0	52706.868657	90928.786678	1000.0	3000.0	25000.0	70000.0	1000000.0
Doctoral degree	2217.0	70641.181777	117160.947589	1000.0	4000.0	40000.0	90000.0	1000000.0

Table 3.0

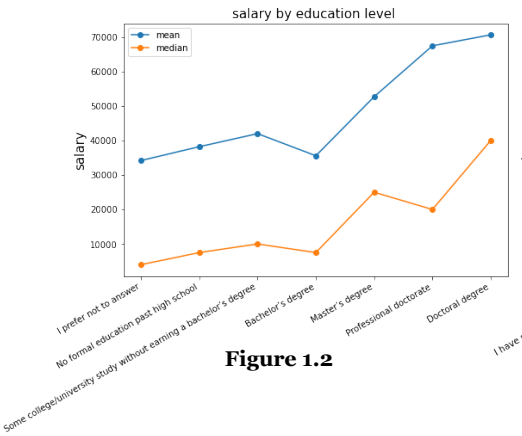

Figure 1.2

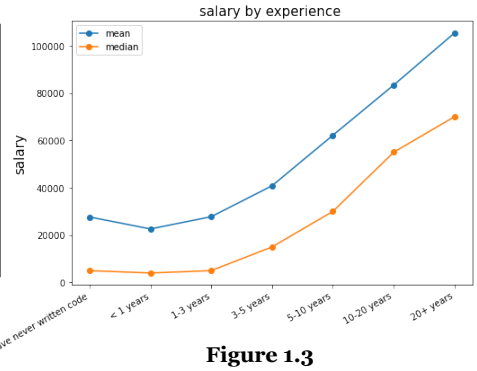

Figure 1.3

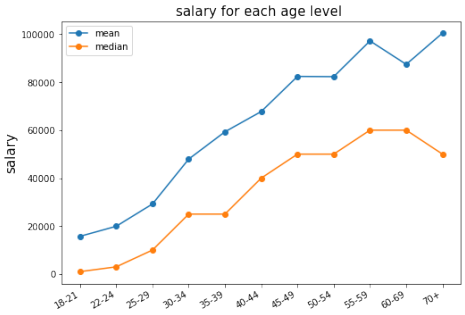

Figure 1.4

Figure 2.1

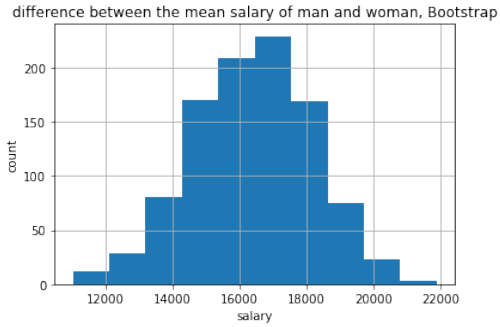

Figure 2.2

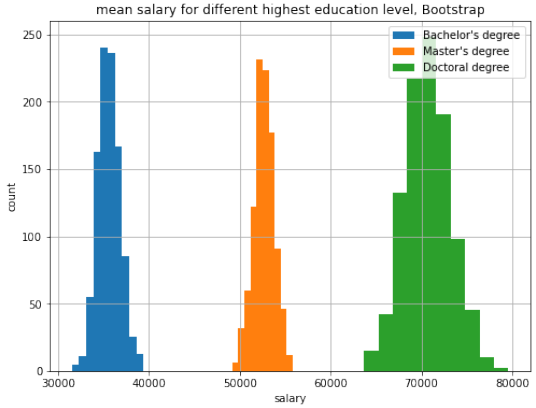

Figure 3.1

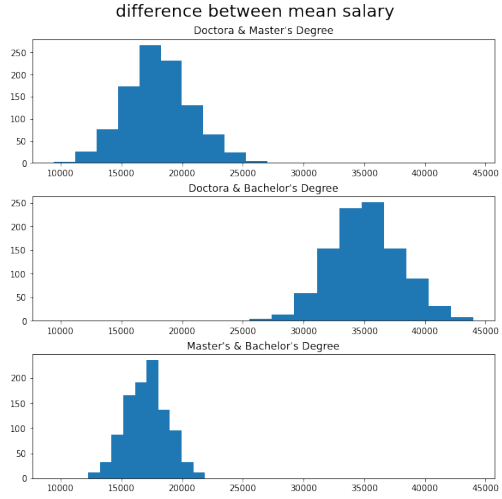

Figure 3.2